STUDENT PERFORMANCE AS A MEASURE OF INSTRUCTIONAL PROFICIENCY

PROJECT NO. NM 001 077.01.06

RESEARCH REPORT

OF THE

U.S. NAVAL SCHOOL OF AVIATION MEDICINE

NAVAL AIR STATION

PENSACOLA FLORIDA

U. S. Naval School of Aviation Medicine 25 June 1954

Project No. MM 001 077.01.06

"Student performance as a measure of instructional proficiency."

Wilse B. Webb, The Tulane University and Norman D. Bowers, U. S. Naval School of Aviation Medicine.

7 pp. 3 tables UNCLASSIFIED

A fundamental question about instructors and instructional techniques is whether there are significant differences between instructors in terms of student achievement. This study is concerned with this question. A critical test of the effectiveness of a flight instructor was secured by determination of his students' performance as flyers. Performance of randomized students of various instructors was found to be significantly different. Therefore, some instructors are more effective than others in teaching students to fly.

1. Instructors - Performance 2. Military personnel - Training
3. U. S. Naval School of Aviation Medicine

I. Webb, Wilse B.
II. Bowers, Norman D.

(L. C. Subj. Head)

U. S. Naval School of Aviation Medicine 25 June 1954

Project No. MM 001 077.01.06

"Student performance as a measure of instructional proficiency."

Wilse B. Webb, The Tulane University and Norman D. Bowers, U. S. Naval School of Aviation Medicine.

7 pp. 3 tables UNCLASSIFIED

A fundamental question about instructors and instructional techniques is whether there are significant differences between instructors in terms of student achievement. This study is concerned with this question. A critical test of the effectiveness of a flight instructor was secured by determination of his students' performance as flyers. Performance of randomized students of various instructors was found to be significantly different. Therefore, some instructors are more effective than others in teaching students to fly.

1. Instructors - Performance 2. Military personnel - Training
3. U. S. Naval School of Aviation Medicine

I. Webb, Wilse B.
II. Bowers, Norman D.

(L. C. Subj. Head)

U. S. Naval School of Aviation Medicine 25 June 1954

Project No. MM 001 077.01.06

"Student performance as a measure of instructional proficiency."

Wilse B. Webb, The Tulane University and Norman D. Bowers, U. S. Naval School of Aviation Medicine.

7 pp. 3 tables UNCLASSIFIED

A fundamental question about instructors and instructional techniques is whether there are significant differences between instructors in terms of student achievement. This study is concerned with this question. A critical test of the effectiveness of a flight instructor was secured by determination of his students' performance as flyers. Performance of randomized students of various instructors was found to be significantly different. Therefore, some instructors are more effective than others in teaching students to fly.

1. Instructors - Performance 2. Military personnel - Training
3. U. S. Naval School of Aviation Medicine

I. Webb, Wilse B.
II. Bowers, Norman D.

(L. C. Subj. Head)

U. S. NAVAL SCHOOL OF AVIATION MEDICINE
NAVAL AIR STATION
PENSACOLA, FLORIDA

JOINT PROJECT REPORT NUMBER 6

The Tulane University of Louisiana
Under ONR Project NR 154-098

and

U. S. Naval School of Aviation Medicine
Research Project Number NM 001 077.01.06

STUDENT PERFORMANCE AS A MEASURE OF INSTRUCTIONAL PROFICIENCY

Report by

Wilse B. Webb, Ph.D.

and

Lieutenant (jg) Norman D. Bowers, MSC, USNR

Approved by

Doctors Cecil W. Mann and Wilse B. Webb
The Tulane University

and

Captain Ashton Graybiel, MC, USN
Director of Research
U. S. Naval School of Aviation Medicine

Released by

Captain James L. Holland, MC, USN
Commanding Officer

25 June 1954

SUMMARY

A fundamental question about instructors and instructional techniques is whether there are enough differences between instructors to require our concern. It is conceivable that such differences are quite minimal and have little effect due to the similar training and backgrounds of the instructors, identical syllabi, and standardization procedures. If such is the case, elaborate instructor training, selection, or monitoring is unnecessary. If there are considerable differences among instructors in instructional proficiency, action is required for efficiency in training. This study concerns itself with the question of the extent of differences between instructors.

Because students are not likely to be completely adequate judges of instructors and supervisors are not likely to have sufficient opportunities for observation of instruction, a measure for instructor differences was developed and used. We have judged instructor differences on the basis of how well an individual instructor's students compare with the students of other instructors. The comparison was made on what should be the critical test of a flight instructor -- how well his students performed as flyers.

The flight performance of the students that 12 officers had instructed through the first 11 flights was compared on the grades assigned the students on check flights (their 12th through 19th flight) and their second stage of training.

It was found that the performance of students of different instructors was significantly different. This could not be attributed to differences between students assigned to the instructors.

Our data indicate that some instructors are not doing as effective a job of teaching students to fly as others. It follows that some cadets are being unfairly penalized by this fact.

Such instructor differences can be minimized by instructor selection, instructor training, standardization and monitoring of procedures, or instructor elimination. It is quite likely no one of these procedures will be successful by itself, but efforts along all of these avenues can effectively eliminate these differences.

TECHNICAL SUPPORTING DATA.....

INTRODUCTION

Perhaps one of the more difficult problems in research dealing with the evaluation of instructor effectiveness is showing or "proving" a significant difference between students of various instructors as a result of received instruction. Nearly all the classically used measures are suspect. For example, student ratings may be questioned on the grounds of the possible incompetence of the student to judge due to lack of experience with the subject matter or bias on such "irrelevancies" as the personal appearance of the instructor. Frequently, supervisor ratings can be questioned on the ground of limited observation opportunities of the supervisor. Although instructor self-evaluations have been used, they lack a needed objectivity.

There seems to be one alternative technique which may resolve this dilemma. This would be an evaluation of instructional proficiency through student achievement. Such a measure would have a number of advantages, the most important being its validity. After all, the most important result of teaching is not how well the instructor is "liked" by his students, his superiors, or himself. Of more importance is the effect his teaching has on the subsequent performance of the learner.

In the final analysis, the most important criterion of instructor effectiveness is student achievement -- including attitudes and understanding as well as skills. Although stated in a variety of different ways, for many years, professional educators have agreed upon the criterion of student achievement as a measure of instructor successfulness. The recent report of American Educational Research Association's Committee on Criteria of Teacher Effectiveness indicated the criterion of teachers' effect on pupil achievement was one of the more ultimate criteria of the "ultimate-proximate" continuum (Review of Educational Research, 22: 238-263, June, 1951).

It is probable that the failure to use student performance as a measure of instructional proficiency results from difficulties of measurement rather than the logic of the technique. For example, a final examination is constructed by an instructor or supervisor and administered to a class which is below average in aptitude or original interests. Scores are then compared with an examination administered to students of a different instructor. Such an analysis would not be particularly revealing. Our reasoning in the above example would be as follows: if the differences among the students of the instructors were not controlled, individual student differences would be measured rather than instructional differences. Further, if the instructor or the supervisor constructed or administered the final examination, this measure would probably reflect what the instructor or the supervisor thought should be taught rather than what the student actually learned.

The best type of measure would necessarily be one which objectively measured performance directly related to the task to be learned, and was independent of unique student, instructor, or supervisor differences concerning what had been or should have been learned. These conditions must

be met if these measures are to be used to compare differences among the instructors in the ability to teach students in a common task.

Recently, the investigators had an opportunity to obtain measures of student performance which seemed to meet most of the criteria required to employ this measure as a measure of instructional differences. This paper is a report of the use of this type of measure.

PURPOSE

The purpose of this report is to describe a test of the null hypothesis concerning instructor differences as they affect student performance in a complex learning situation. Stated in another fashion: there is no difference between instructors in their effect on student learning in a complex learning situation. If differences do exist, we would be in a position of saying that instructional differences play a role in the efficiency of learning. If the null hypothesis is found tenable, our attention must be directed then to possibilities that: (a) instructor differences in the particular situation were minimal; (b) student learning in this particular situation was essentially independent of instructor technique; (c) instructors so effectively distribute their training that instructor differences were compensated for; or (d) perhaps the measure was not sensitive enough to reflect differences in student performance.

PROCEDURE

This study was carried out in the Naval Air Training Command. The complex task to be learned was that of flying an aircraft. Student subjects were Naval Aviation Cadets; the instructors were Naval officers assigned to the Command for instructional duty. Although the instructors had extensive training and experience in flying, as a group they had very limited education in methods of instruction.

The period covered by this investigation was between approximately February 1952 and June 1953. During this period, the records of the students trained at one of the initial training fields were sorted on the basis of their instructor. Then, the students who had had the same instructor through the first 11 flights were selected. The final data for this study were the student records of 12 instructors who had four or more students during the period considered.

The students were given instructions for 11 flights by their instructors. This is a phase of training during which the student is grounded on the basic fundamentals of flight (landings, take-offs, turns, taxiing, etc.). At the end of this period, he is then given a "check flight" by another instructor. This is designated as A-12 flight. After successful completion, the student receives six more dual flights, usually accompanied by different instructors. He is then given a further independent "check flight" by an instructor other than the one he has previously had. This is the A-19 flight. After his solo, he then receives one or two more flights and proceeds to his "B stage" of training. This amounts to some 17 flights in

which he "practices" previously learned techniques. This phase is known as the precision flying training. In addition to "practicing" the fundamentals he has learned, he also learns to recover from stalls, to make small field landings, and so on.

The progress described above is somewhat idealized in the sense that "extra time" is allotted individuals who are not proficient in various techniques at various stages. For example, if the man is not considered safe to solo on the A-19 flight, and he is considered potentially capable, he may be given several extra flights, re-checked, and permitted to solo. A man, however, can be dropped if found considerably wanting at any given stage.

The grades of the A-12 flight, the A-19, and the grades made during the "B Stage" of training were used as criteria. As indicated, the grades given during the A-12 and the A-19 flights are essentially progress checks to evaluate student learning. In a sense, the "B Stage" is a practice period and an extension of learning during the early stage of training. As was indicated, these grades are assigned independently of the original instructor. They should serve as an evaluation of the proficiency of the individual, independent of the instructor.

The grades are averages of the ratings by instructors on various maneuvers on which the student has had training. If the performance of a particular phase (e.g., landings) is unsatisfactory, then the student is given a grade of 1.0 on that maneuver. If below average, a grade of 2.0 is given, if average, 3.0, and if above average, the grade of 4.0 is assigned on that maneuver. The grade for a given flight is the sum of these separate ratings divided by the number of maneuvers which were rated. The average of the grades is approximately 3.0, and a grade of 2.5 for a given flight is considered unsatisfactory and the student is either dropped from training or given extra time. These grades should reflect how well the individual has learned to perform the task for which he has been instructed; namely, efficiently flying an aircraft.

In several cases, a student of an instructor attrited before the "B Stage." In these cases, he was assigned a grade of 2.5 (considered unsatisfactory), and included in the determination of the student record of the instructor.

An analysis of variance was performed on these grades by instructors. A second analysis of variance of the aptitude scores of the students of each instructor was performed to check the assumption that no initial differences existed among the students of the various instructors.

A brief summary of the preceding description of our procedure is as follows:

a. Instructors - 12 experienced Naval Aviators.
b. Students - 4, 5, or 6 trainees, without previous flight experience, assigned to each instructor.

c. Instruction - 11 training flights of approximately one hour and twenty-five minutes each in which the instructor attempted to give each student the basic fundamentals of flying an aircraft.

d. Measures - Evaluations of student ability to fly a Naval aircraft immediately after the 11 initial flights (the A-12 flight); a further progress check of the student's flight proficiency after 18 flights (the A-19 flight); the average performance in a second stage of training in which the student learns to become more precise and detailed in his flying techniques (B Stage). This latter measure consists of some 21 flights which the student performs after having had 24 previous training missions. Each of the measures used were given by instructors other than the original instructor.

e. Controls - It was assumed that no individual differences existed between the students of the various instructors. This assumption was tested and reported in our results. Further, it is reasonable that the measures taken are independent of any unique instructor or supervisor bias in regard to what the student should have learned or had learned, since the grading system has been developed independent of the instructors and their immediate supervisors. Finally, it can be easily maintained that the measures were related to the task to be learned.

RESULTS

The means of the student flight grades on the two check flights (A-12 and A-19) and the latter stage of training ("B Stage") for the 12 instructors are given in Table I. Further, the means of the Flight Aptitude Rating scores, a measure of flight aptitude, are given in Table I. The table also shows the number of students on which each measure was based.

An F max test of the homogeneity of variance for the three check hops was completed. There was no significant difference in the variance of the three groups, i.e., it may be concluded that the variances are homogeneous. A probit analysis of the three sets of ratings was performed and indicated that the distributions of check ride scores are approximately normal. An analysis of variance between the students' Flight Aptitude Rating* scores was completed, and is given in Table II. There was no statistically significant difference between students. The FAR is predictive of success in the Flight Training Program, and this information can be interpreted to

---

* Hereafter referred to as FAR. It is a composite score based on three different measures purporting to predict success in the FT program.

mean that, regarding the predicted ability to complete the program successfully, the students of one instructor did not differ from the students of another instructor. Therefore, any differences which are found in final achievement must be explained as resulting from some effect other than ability to complete the Flight Training Program. The F max test of homogeneity of variance, the probit analysis, and the analysis of variance of FAR scores represent tests of assumptions underlying the analysis of variance technique. On the basis of this information, it was concluded that further analysis of data would be justified.

Accordingly, an analysis of variance was performed between the A-12, A-18, and B stage check ride grades for 62 students who had been assigned to 12 different instructors. Because of the unequal frequencies in the cells, Tsao's method of solution was utilized for the resulting 4 x 12 design (Psychometrika, 11:107-128, June, 1946).

The results of this analysis are shown in Table III. The interaction term (student achievement x instructors) was not significant. Therefore, it was pooled with the "within variance" for testing the significance of the main effects (variation between instructors, and variation between student achievement). Differences significant at the .01 level were found between instructors and between student achievement on the three check rides. The results of this study indicate that there is a difference between instructors regarding the achievement of their students, and there is a difference between students in achievement from check ride to check ride.

CONCLUSIONS AND DISCUSSION

It will be recalled that the null hypothesis tested in this investigation was concerned with instructor differences as they affect student performance. Measures were obtained of student performance which were independent of the evaluation of the student by the instructor or the instructor by the student. It was found that significant differences existed between the performances of different instructors, and the null hypothesis was rejected. It is clear that the differences between instructors significantly affected performance of students.

Our finding may be little more than a proof of something we all agree on. We hope, however, that these data will serve as a twofold reminder: (1) that instructor differences do result in student achievement differences, and (2) that methodologically these differences can be established independent of the arguments surrounding the other techniques of evaluating instructor effectiveness.

The example of complex learning (aircraft flying) might be considered a special case in that the instructor-student relationship was on a face to face relationship, rather than in a classroom group situation. However, the flight student has had no previous experience or training in flying aircraft. It is one of the few learning situations which is available which lends itself to the evaluation of the effects of one instructor. Such a

situation has some import in other educational situations. Basically, the teaching of the basic skills entails a teacher-student relationship which is quite independent of group effects. One has but to look at the wealth of educational literature to see the importance attached to the "individualization of instruction," "teaching for individual differences," and "promoting individual attention." However, the teaching usually referred to does take place in a group situation, and it is difficult to ascertain accurately the effects of the instructor. Further, after an initial learning experience in one of the basic skill subjects, the effects of later instruction are impossible to ascertain. In our opinion, this particular study offers definitive evidence of the effects of instructors, and gives objective evidence to the contention that different instructors do have a differential effect on student learning. This principle, it is felt, can be generalized to learning situations in which basic skills are taught. As such, this study represents evidence regarding the importance of the instructor, a concern for instructor methods, and thereby a concern for instructor education.

TABLE I

FLIGHT APTITUDE RATINGS AND GRADES FOR THE A-12, A-19, AND B STAGE FLIGHTS FOR 62 STUDENTS ASSIGNED TO 12 INSTRUCTORS

Instructor	A				B				C			
Student	1	2	3	4	1	2	3	4	1	2	3	4
FAR	5	9	7	7	3	9	7	5	5	6	8	5
A-12	2.76	3.00	2.71	3.07	2.53	2.79	3.17	2.53	3.00	2.50	2.82	2.50
A-19	2.50	3.00	2.89	2.78	2.67	2.89	2.58	2.42	2.50	2.50	2.95	2.50
B	2.50	3.10	2.95	2.81	2.87	2.85	3.01	2.92	2.50	2.50	2.75	2.50

Instructor	D					E					F				
Student	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
FAR	7	8	6	6	9	4	6	8	8	8	5	7	6	9	9
A-12	3.00	3.00	2.50	2.94	3.00	3.06	3.17	2.63	2.65	3.00	2.72	3.11	3.67	2.83	2.35
A-19	2.50	2.53	2.50	2.81	3.16	2.50	2.42	2.89	2.89	2.44	2.71	2.83	3.16	3.06	2.68
B	2.50	3.14	2.50	2.94	3.06	3.06	2.69	3.07	3.01	2.94	3.05	3.13	3.24	3.02	3.02

TABLE I (Cont'd)

Instructor	G				H					I				
Student	1	2	3	4	1	2	3	4	5	1	2	3	4	5
FAR	9	8	5	5	5	6	5	5	8	9	5	6	3	5
A-12	2.94	3.00	2.72	2.50	3.05	2.89	2.82	2.83	3.00	2.94	2.94	2.51	2.81	2.81
A-19	2.21	2.79	2.58	2.50	2.80	2.47	3.05	2.19	2.84	3.11	2.84	2.83	2.89	2.95
B	2.95	3.02	2.50	2.50	2.91	3.03	2.75	2.97	2.94	2.83	2.81	2.80	3.10	3.22

Instructor	J								K					
Student	1	2	3	4	5	6	7	8	1	2	3	4	5	6
FAR	5	4	3	7	6	8	8	6	9	6	9	4	6	8
A-12	2.83	2.50	3.00	2.75	2.50	3.24	3.00	2.50	2.67	2.83	2.88	2.50	2.63	2.94
A-19	2.79	2.50	2.84	2.84	2.50	3.26	2.93	2.50	2.95	2.89	3.00	2.50	3.26	3.00
B	2.75	2.50	2.89	2.85	2.50	3.02	2.97	2.50	2.98	2.92	3.03	2.50	3.03	3.03

Instructor	L						
Student	1	2	3	4	5	6	7
FAR	6	7	7	9	7	6	6
A-12	3.00	2.93	2.94	3.00	2.78	2.94	3.06
A-19	2.89	2.50	3.00	2.22	2.33	2.84	3.33
B	3.03	2.91	2.83	3.00	2.94	2.86	3.14

TABLE II

ANALYSIS OF VARIANCE OF FLIGHT APTITUDE RATINGS
OF 62 STUDENTS ASSIGNED TO 12 INSTRUCTORS

Source	df	Sum of squares	Mean square	F
Between instructors	11	30.6178	2.78	< 1
Within groups	50	144.8822	2.90	
Total	61			

TABLE III

ANALYSIS OF VARIANCE OF 62 STUDENTS ASSIGNED TO 12 INSTRUCTORS ON THREE SETS OF CHECK FLIGHT GRADES: A-12, A-19 AND B STAGE

Source	df	Sum of Squares	Mean Square	F
Between flights	2	1.957059	.9785	18.53**
Between instructors	11	3.064799	.2786	5.28**
Residual	172	9.074092	.0528	
Interaction	22	.861892	.0392	< 1
Error	150	8.2122	.0547	
Total	185			

** Significant at the .01 level.